Am J Hum Genet 33:664–682, 1981

The William Allan Memorial Award Address: Gene Clusters, Genome Organization, and Complex Phenotypes. When the Sequence Is Known, What Will It Mean?

WALTER F. BODMER[1]

I am very conscious of the honor that the American Society of Human Genetics has bestowed upon me by choosing me to receive the Allan Award in 1980. It is, I think, the only major award specifically for human genetics, and its international flavor is documented by the fact that I, a Britisher, although admittedly a renegade immigrant to the United States, have been chosen. In this address, I have the responsibility of following a distinguished line of predecessors who reflect all aspects of our ever-expanding and increasingly important field of human genetics.

When I was in high school, I remember often trying to match my friends to their parents at various school functions and being surprised at how easy this was. As human geneticists, in spite of the enormous advances being made in our field, we still cannot answer many of the everyday questions that we are asked, such as: "Why does he look just like his mother?" Max Perutz [1], in a recent editorial comment in the *New Scientist* entitled "The Molecular Biology of the Future," suggested some questions, for, as he put it, "an examination in some future century." Here are two of them: (1)

> "The time has come" the Walrus said,
> "To talk of many things . . .
> And why the sea is boiling hot
> And whether *pigs have wings*."

Calculate the amount of genetic information this would require in megacricks. The *Genetic Tables for the Design of Domestic Animals* may be used for reference. (2) Prescribe a therapy at the molecular level for Hamlet.

Will it be a future century before we can answer such questions? Recent advances in molecular genetics have surely made it likely that that "future century" will be the next century, and "the time," perhaps, within the lifetime of many contemporary human geneticists, at least for some of the answers. The aim of this paper is to outline some of the prospects and the presently available approaches as

---

Presented on the occasion of receiving the William Allan Memorial Award at the annual meeting of the American Society of Human Genetics, New York, September 24–27, 1980.

[1] Imperial Cancer Research Fund, Lincoln's Inn Fields, London, England.

I see them, via a brief tour of the HLA system and the overall organization of the genome.

My entry into genetics as a mathematician and statistician working under R. A. Fisher was from the top in terms of complexity. It involved the statistical analysis of populations, of complex phenotypes, and of their patterns of inheritance. Soon I came to realize the importance of descending from the top to the concrete realities of molecular biology; hence, my transition from Fisher to Lederberg, from the top in terms of complexity, down to the more analyzable molecular level. Now, of course, the challenge is to climb from the bottom up and to use the knowledge of molecular genetics and recombinant DNA techniques to go from the molecular bottom up again to the analysis of the complex phenotype.

No journey is ever straightforward, and so in my descent to the fascinating depths of molecular genetics at Stanford, I was waylaid through my statistical background, and remembering Fisher's incisive analysis of the Rhesus system, into studies in collaboration with Rose Payne and Julia Bodmer that contributed to the development of the HLA system. This has become for me a paradigm of the complex gene cluster, or supergene, which, I believe, reflects, at the upper level of complexity, the general pattern of organization of the genome of higher eukaryotes.

THE HLA SYSTEM

The HLA system was first defined as a series of antigenic specificities identified by using sera produced by fetal-maternal stimulation on peripheral white blood cells, especially lymphocytes, in a search for white blood cell groups that could form a basis for transplantation matching. The HLA region and its counterpart in the mouse H-2 are known to include a number of closely linked loci controlling cell surface determinants as well as genes for immune responsiveness and certain components of the complement system (see, e.g., [2]). These regions encompass a recombination fraction of between 1% and 3%, and so certainly qualify as supergenes as defined by Darlington and Mather [3] to be "a group of linked genes mechanically held together on a chromosome and usually held together as a unit."

A schematic comparison of the H-2 and HLA genetic maps in relation to the linked glyoxylase (GLO) marker and the centromere is shown in figure 1. The *HLA-A*, *B*, and *C* loci code for the originally defined antigens, which are present on the majority of tissues. The *HLA-D* locus codes for the determinants defined by the mixed lymphocyte culture reaction, while *DR* codes for the corresponding serological specificities found mainly on B lymphocytes and monocytes or macrophages. The second and fourth complement components and factor B, the analog of C-2 in the alternate complement pathway, are coded for by genes between *HLA-B* and *D*. In the mouse, the Ss protein is actually C4, while the human Chido and Rogers and mouse G red cell blood groups correspond to a portion of C4 attached to the surface of the red cell. Thus it seems most probable that these complement coding regions in mouse and man are homologous. In the mouse, *H-2K*, *D*, and *L* correspond to *HLA-A*, *B*, and *C*, the I region, defined originally by its control of the immune response, includes the mouse equivalent of *HLA-D*(*R*), while the human homologs

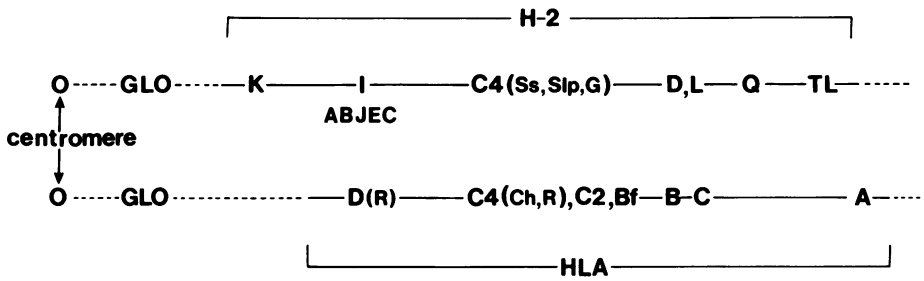

FIG. 1.—Schematic comparison of the genetic maps of the mouse H-2 and human HLA regions. *Ch, R* refer to the Chido and Rogers blood groups, *GLO* to the enzyme glyoxylase.

of the mouse *H-2Q* and *TL* loci, which control further sets of lymphocyte antigens, have yet to be defined. The orientation of the human and mouse maps shown in figure 1 aligns five out of six of the genetic markers identified in both species, counting the centromere, and so leaving only *H-2K* and *HLA-A* out of line. The position of the human equivalents of *Q* and *TL* in relation to *HLA-A* and *C* remains to be established. There are various ways in which the two sequences can be derived from each other; for example, by two inversions, or by an intrachromosomal unequal double crossover [4].

The molecules recognized by antisera to the HLA-A, B and C products are composed of a 43,000-mol. wt. glycosylated polypeptide that carries the polymorphic specificities and that is noncovalently linked to $\beta 2$ microglobulin, a 12,000-mol. wt., nonglycosylated polypeptide. While the HLA region is on chromosome 6, the gene for $\beta 2$ microglobulin is on chromosome 15. Molecular data clearly suggest that the *HLA-A*, *B* and *C* loci are related to each other by gene duplication. The molecules on the cell surface recognized by HLA-DR antisera comprise two noncovalently associated glycosylated polypeptides with mol. wts. of approximately 33,000 and 28,000, the latter carrying the polymorphic determinants [5–7]. There is, in addition, a third chain of mol. wt. 30,000, first found in the mouse by Jones et al. [8], which associates with the other two chains, although apparently not on the cell surface and under circumstances that still remain to be clarified. In the mouse, the I-A and I-E/C products, although they are similar to each other, differ substantially, and it appears that the I-E/C product is the homolog of the HLA-DR product as identified so far. However, there is evidence for heterogeneity of HLA-DR products, and these may either correspond to the products of two very closely linked loci, with the human equivalent of I-A still to be identified, or simply to I-E/C and I-A, or to both [4]. Data in the mouse suggest that the I region controls the production of both the 28,000- and 33,000-mol. wt. products, and this is usually interpreted to suggest that the structural genes for both sets of products are in the I region. However, as pointed out by Barnstable et al. [9], this would be somewhat unusual, as in other cases in which polymeric proteins are made up of different subunits, the genes for these are unlinked. Exceptions such as insulin and C4 arise when a single large precursor polypeptide is processed to form two or more separable subunits, but there is evidence against this possibility for the HLA region

products [10]. This emphasizes the distinct possibility that the structural genes for the 33,000-mol. wt. chains may not be coded for in the H-2 or HLA regions. The genetic control exerted by the H-2 and HLA regions may be with respect to which of a set of alternative chains coded for elsewhere in the genome is expressed. Conventional genetic crosses do not, in fact, distinguish between control of expression by a gene and coding for the gene product. For this distinction to be made unequivocally, either somatic cell hybrids or direct molecular localization using DNA techniques are needed.

The complement components C2 and factor B have similar structures, each being single polypeptide chains with an approximate mol. wt. of 100,000, while C4 has a mol. wt. of 200,000 and is comprised of three polypeptide chains derived from a single large precursor molecule.

The HLA-A, B, and C products are required for immune T lymphocyte recognition of specific target antigens, while the HLA-DR products are involved in interactions between T and B lymphocytes and between T lymphocytes and macrophages. Thus, the major identified functions of these products are to play a role in cellular interactions in the immune system. One interpretation of this function, which provides a natural rationale for associating genes for complement components with those for the HLA-A, B, C, and DR products in the same genetic region, is that the latter may behave like complement components but act on the cell surface in relation to the immunoglobulin-like T cell receptor. In this way, the antibody-antigen-complement triple interaction is assumed to have its parallel in the T receptor-antigen-HLA region product association on the cell surface. These functions can provide a direct explanation for the genetic control of immune response differences by the H-2 and HLA regions [11].

The main evidence in man for genetic control of immune response by the HLA region comes from a number of striking associations between particular HLA determinants and a variety of chronic diseases that are either autoimmune or have some form of immunological etiology. This immune response control, especially if it underlies differences in susceptibility to infectious diseases, can easily form a basis for an explanation for the extraordinarily high level of polymorphism observed for the serologically detected HLA determinants ([12, 13]; see also, e.g., [14] for further background on the HLA system).

There has been much speculation as to the possible role of HLA determinants and their H-2 counterparts in cellular interactions outside the immune system (see, e.g., [12, 15]). The expression of HLA-DR determinants on melanomas and on certain epithelial tissues that have nothing to do with the immune system [6] and the association of HLA-A3 with hemochromatosis and Cw6 with psoriasis, neither of which are likely to involve any aspects of the immune system, fit in with this possibility. But, so far, there is admittedly no conclusive experimental evidence for functions outside the immune system.

The present view of the HLA system shows it, therefore, to be a complex genetic region coding for a variety of different protein products that occur in sets clearly related to each other by duplication and controlling functions that are to a considerable extent interrelated, involving the cell surface and the immune system. There

is, however, one potential puzzle. The 21-hydroxylase deficiency gene has been clearly shown to be very closely linked to the HLA region, and some data suggest that it may be within the region [16]. If this were the case, then surely one would have to ask what such an apparently alien gene was doing in the midst of all the other genes controlling surface and immune-related functions.

HLA IN PIECES

Genes were blown to pieces in 1977! No longer did the simple one to one relationship between DNA sequence and corresponding product exist. Now we know that coding sequences are interrupted by intervening noncoding sequences, that a protein product may be made up of many different coding regions corresponding, perhaps, to its different domains, and that a given coding region may contribute to more than one gene product, as for example in the case of the immunoglobulins (see, e.g., [17] for review). These ideas, strangely foreshadowed by Pontecorvo in 1959 [18], but certainly not anticipated by most, in fact, provide the basis for much more satisfactory explanations of some observations on the HLA products. They also provide an opportunity for rampant speculation before the facts are established by the DNA sequence. As Julia Bodmer put it in her "*Typer's Lament*," written for the 8th International Histocompatibility Testing Workshop,

As all the small pieces were dropped on the floor,
I picked up ten pieces and you picked ten more,
Now some are unique and some overlap—
Looks like many loci—but it may be a trap!

Established serological patterns of cross-reactivity, new complexities revealed by monoclonal antibodies [19], and data on the chemical structure of the HLA-A, B, and C determinants [20] and their H-2 counterparts [21], all suggest that these products can be subdivided into molecular regions corresponding to distinguishable domains, which, in turn, are likely to correspond to noncontiguous DNA coding regions, as described in many other systems. At the protein structure level, the domains incorporate disulphide loops, membrane-embedded hydrophobic regions, and variable regions, likely to be those that determine the polymorphic differences. A particularly interesting feature of the amino acid sequence is the homology of one of the domains with one of the immunoglobulin IgG heavy chain constant region domains, analogous to the similar well-established homology between $\beta 2$ microglobulin and immunoglobulins. This recalls earlier speculation concerning the relationship between HLA and immunoglobulin products. The relationship between serological domains and the amino acid sequence remains to be established. A particular feature, however, of the data using monomorphic monoclonal antibodies to the HLA-A, B, and C, products is the definition of one or more antigenic sites common to these three products. Evolutionary considerations suggest that these sites are likely to be coded for by a common nucleotide sequence. Such a region has also been proposed by Demant et al. [22] in the H-2 system to account for single polymorphic determinants found on all three of the H-2 K, D, and L products.

A highly schematic view of a part of the HLA region coding for the HLA-A, B, and C products is shown in figure 2. The "A, B, C common" region of the DNA codes for a piece common to the HLA-A, B, and C products carrying the monomorphic determinants recognized by the monoclonal antibodies. Further regions are assumed to code for domains of the protein product that carry locus-specific, cross-reacting, and allele-specific determinants. Polymorphism for control of gene expression [23] now becomes polymorphism for which DNA region of a set coding for a particular domain of the protein is used in the construction of a given product.

This picture poses more questions than it seeks to answer. Is the grouping of the regions by product rather than by type of region correct? Given that the recombination fraction between *HLA-A* and *HLA-B* is of the order of .08%, and that this is likely to represent a distance of at least hundreds of thousands of nucleotides, how are the pieces put together over such relatively long distances? Is it possible that regions of the DNA loop out, so that transcripts can be made directly from noncontiguous DNA sequences, the loops being held in place by small RNAs as suggested for the control of splicing by Steitz, and her colleagues [24] and by others [25]? If these small RNAs are coded for well outside the HLA region, does this provide a mechanism for control of expression of products by unlinked genes, as may be the case for one of the constituent polypeptides of the HLA-DR product? What might be the nature of the signals that control which of a multiple set of alternative regions is expressed by any given chromosome?

CLUSTER EVOLUTION

The significance of gene duplication for the evolution of new gene functions has been discussed since Bridges's original definition of the phenomenon [26] and has formed the basis for suggestions of how supergenes, or gene clusters as I prefer to call them, might have evolved [27, 28]. The simple theory is that once a gene has been duplicated, the copies can diverge from the original and so acquire new functions while not jeopardizing those functions fulfilled by the original genes. The basic and most intensively studied molecular model for such a mechanism is the hemoglobin β cluster (see, e.g., [29]). The problem posed for such a model by the HLA system is to explain how duplication of a single primordial nucleotide sequence could give rise to several apparently structurally unrelated polypeptide products.

The new look of genes in pieces at the molecular level provides a number of possible answers and emphasizes the complex ways in which gene clusters may evolve. The minimal functional unit now becomes the "structural" DNA sequence that, uninterrupted by an intervening sequence, codes for a protein domain. As Crick [17] put it, shuffling around such structures is a convenient way of combining properties of parts of various proteins into a new protein. Obvious examples might be the signal peptides found at the end of newly synthesized membrane and secreted proteins, the active sites of certain enzymes, particular functional regions of a protein such as that embracing the hem group in hemoglobin, and regions with

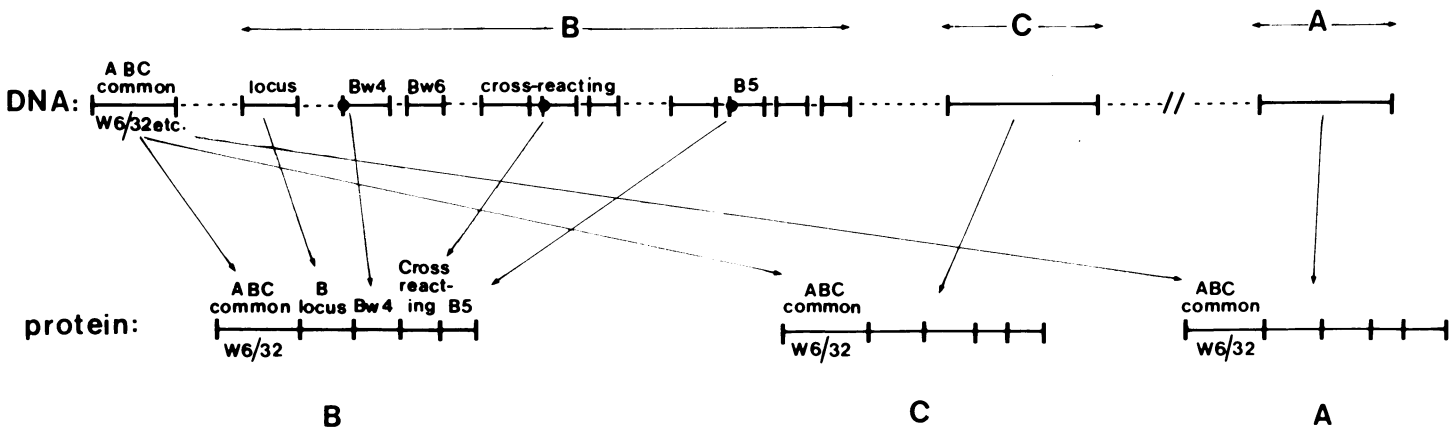

FIG. 2.—Scheme for the production of HLA-A, B, C products from different DNA regions corresponding to different domains. *Continuous lines* in the DNA sequence correspond to coding regions that may themselves contain intervening sequences. *Dotted lines* are flanking regions. *Large dots* at the *left* of a region such as Bw4 indicate that this is the one of a set of alternatives that is being expressed. It is assumed that the derivation of the A and C protein products is similar to that for B. *Arrows* from DNA to protein indicate which DNA region is coding for a given protein domain. // between A and C indicates that other regions, perhaps corresponding to Q and TL, lie there.

particular structural features such as the collagen-like sequence found in the complement component C1q [30]. Shuffling can, of course, occur within a gene cluster, and different products may be formed from different combinations of domains within a cluster. Domains may be duplicated singly or in groups together with their intervening sequences and flanking regions, and subsequently diverge by normal evolutionary processes, or they may be transposed by conventional recombination mechanisms or by less conventional means. DNA sequences can be read in six ways, three reading frames in each direction, so that small changes at the nucleotide level can give rise to quite new products [31]. An interesting example of this is the suggestion by Fiddes and Goodman [32] that the β subunit of human chorionic gonadotrophin might have evolved at one end by a "read through" into a previously untranslated region.

When a new amino acid sequence is formed that has no obvious relationship to any previous sequence, such as will happen in the case of translation of a frame-shifted stretch of DNA or, as mentioned above, translation of a previously untranslated region, how is that new amino acid sequence put to use? Can a more or less arbitrary amino acid sequence perform some function, however inefficiently, and then gradually evolve to be more efficient? When such a new sequence evolves, can it function in a way that is related to the function of the products already made by the region? Is it possible that in this way the genetic structure moulds the phenotype [33]?

The sort of complex genetic region envisaged in this discussion is likely to provide the opportunity for relatively high rates of spontaneous genetic changes involving complex events. These can include inversions and deletions mediated by intrachromosomal unequal crossing over, excision from the chromosome followed by amplification as an extra chromosomal element and reinsertion into the chromosome, and complex gene conversions in heterozygotes during meiosis as suggested by Nishioka et al. [34] for the evolution of the "dead" third murine α-globinlike gene. Such possibilities are bound to have implications, not only for interpreting the nature of the mutations that lead to some of the simply inherited human syndromes but also for the evolution of intervening sequences and flanking regions.

There are already intriguing data to suggest that the rate of evolution of many noncoding regions is much higher than that of the coding regions [35]. Some would argue that this is expected because these regions are likely to have little functional significance and so random genetic drift will change them at the highest possible rate. It seems to me dangerous, firstly, to assume that these regions will be of little functional significance. Their particular sequence may turn out to be important, in many cases, for aspects of the control of gene expression. But most of all I find it difficult to accept the notion that the fastest possible rate of evolution is for neutral regions, since population genetic theory clearly indicates a slow rate of evolution by random genetic drift. The corollary of this, namely, that there is strong selection acting on intervening sequences and flanking regions, I also find hard to accept, and so we are left with an apparent paradox. How do such regions diverge rapidly during evolution if neither selection leading to rapid change nor neutral random genetic drift are adequate to account for their rate of evolution? Mutation pressure,

even if mutation rates are high, is not an answer since at the molecular level each mutation is essentially a unique event and care must be taken not to confuse high mutation rates, which are still not high enough to be effective evolutionary forces, from rates of incorporation of new mutations into the population. It is accounting for the rapid rate of incorporation into the population that creates the difficulty. One intriguing possibility would be asymmetrical gene conversion in heterozygotes during meiosis. Asymmetrical gene conversion, by which is meant a relatively high rate of production of the homozygote *A1A1* rather than of *A2A2* from the heterozygote *A1A2*, has been observed for certain allelic combinations in fungi [36]. Perhaps heterozygotes for complex mutational events, involving especially insertion or deletion and so, presumably, looping out of the DNA of one parent strand during pairing at meiosis, might lead to preferential excision of material from one or the other parental strand and so in this way to a form of directed meiotic drive for certain sorts of new mutations. Such changes could occur at a fairly high rate, especially if unopposed by natural selection.

THE NUMBER OF CLUSTERS AND GENOME COMPLEXITY

Gene clustering is turning out to be a widespread, almost universal, phenomenon in higher eukaryotes. The clusters may be very complex, as in the case of the HLA system or the immunoglobins, relatively simple, as in the case of the hemoglobins, or, perhaps, even just controlling a single product, as may be the situation for some enzymes. The HLA system appears likely to be at the upper limit of cluster complexity, especially as it involves such a diverse collection of protein products. Even if the ideas discussed above as to how the region might have evolved by duplication from a single, admittedly complex, primordial sequence turn out to be wrong, the region will at least be a cluster of clusters. Widespread and persistent linkage disequilibrium within the region strongly suggests some sort of selective interaction and, so, some form of functional integration within the region as presently defined [13]. This may, perhaps, be mediated through mutually related control of expression of the various products and by functional interaction between them. Within the genetic organization hierarchy, as illustrated in figure 3, gene clusters appear to be the basic functional units. Thus, overall genetic complexity should be considered in terms of the number of gene clusters, rather than in terms of any other genetic unit.

There has been much argument in the past about the number of functional genes in higher organisms. Highly repetitive DNA sequences with no obvious function other than their own "selfish" replication (selfish, that is, in relation to the overall requirements of the organism that harbors them) [37] account for perhaps as much as 50% of the total of 3×10^9 nucleotide pairs in the human haploid genome. The new molecular genetics, however, tells us that of the remaining unique sequences only a small proportion are actually involved in coding for protein. For example, in the hemoglobin cluster, which may be estimated to be about 60,000 base pairs (bp) in length, only five peptides, each containing 146 amino acids, are coded for, which gives a coding ratio of about 1 in 30. In other words, just 1/30 of the total region codes for protein, the remainder corresponding to intervening sequences and flank-

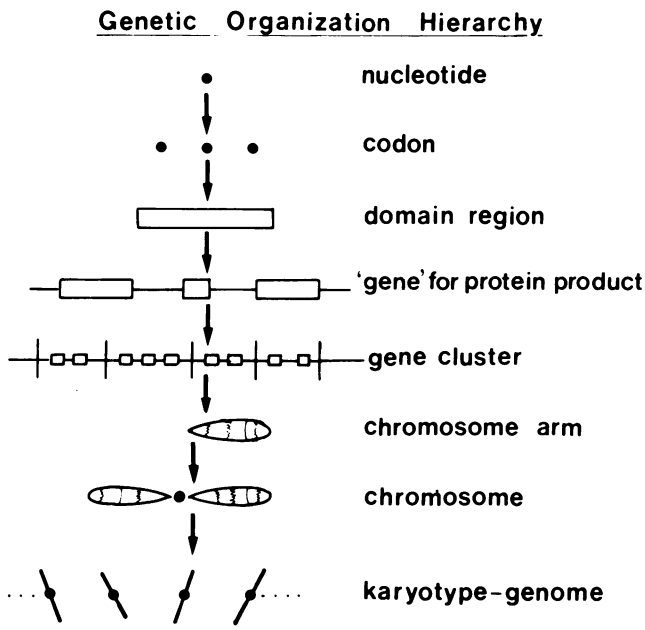

FIG. 3.—Genetic organization hierarchy

ing regions between the genes [38]. The only estimate so far for the length of the HLA region is based on the observed recombination fraction. Assuming the total HLA map length to be three times the *HLA-A* to *-B* interval, namely 2.4 centimorgans, this makes the region about 1/1,000 of the total genome in length, assuming uniformity of recombination throughout the genome. In this case, the region would contain $1/1{,}000 \times 3 \times 10^9$ or approximately 3×10^6 bp. Assuming that the coding ratio of the hemoglobin β region applies, the coding potential for the HLA region would be $1/30 \times 3 \times 10^6 = 1 \times 10^5$ bp, corresponding to about 30,000 amino acids or, for example, about 100 peptides of the approximate size of the HLA-A, B, and C products. Clearly, if these calculations are right, much still remains to be discovered in the region.

What is the likely number of gene clusters in the human genome? This will be a function of (1) the proportion of the total of 3×10^9 bp which are functional, in the sense that they are part of gene clusters rather than selfish DNA; (2) the distribution of the number of genes per cluster, where gene is taken to mean that region of DNA, including intervening and flanking sequences, from which a product is made; (3) the distribution of the sizes of gene products; and (4) the distribution of coding ratios.

There is so far only sparse data on the basis of which to estimate these various distributions, or even to estimate their means. For example, if we assume that 50% of the total DNA is in functional clusters, an average size of gene product of 300 amino acids or approximately 1,000 bp, an average cluster size of 15 products (about the geometric mean between the HLA and hemoglobin clusters) and a coding ratio of 1/30, as in the case of hemoglobin, then the total number of clusters is

$1.5 \times 10^9/30 \times 15 \times 1{,}000$, or approximately 3,300. Clearly, at the present time, this figure is subject to a considerable margin of error, particularly because of the limited data on numbers of genes per cluster and the wide observed variations in coding ratio, from about 1 in 80 for dihydrofolate reductase (R. Schimke, personal communication, 1981) to about 1 in 7 for ovalbumin [39]. Even in these cases, however, the data are not clear-cut because neighboring clusters have not yet been identified, and so the distance between neighboring clusters is not established. As more information becomes available, it will be possible to obtain more precise numbers and, in particular, some idea of their actual distribution and, therefore, the distribution of cluster sizes. However, it seems likely that the total number of clusters will be in the range of, say, 3,000 up to a maximum of 15,000, while the total number of different protein products may be of the order of 50,000 to 100,000. Assuming that, in general, any given cell type will be synthesizing a small proportion of the products of any given gene cluster and that, since differentiation is mediated by differential gene expression, a given cell is likely to express products from a relatively small proportion, say 10%–20% of all gene clusters, the number of protein products made by any given cell is unlikely to be more than a few thousand. This is consistent with the number of products detectable by two-dimensional gel electrophoresis, although, of course, the possibility that there exists a class of products made in very low amounts and so not detectable in this way cannot be ruled out.

Nevertheless, the general implication is that the number of functional classes of gene products is in the thousands. Furthermore, the overall complexity may be even less than this when the organization of clusters into related families, for example, the two hemoglobin clusters or, more distantly, the immunoglobulins, the HLA system, and $\beta 2$ microglobulin, is taken into account. Functional complexity is, it seems, achieved by making use of combinations of combinations.

Given this estimated complexity of the human genome, and assuming that a major element of control of splicing, and gene expression in general, is by small RNAs as suggested by Steitz and others, how many such RNAs might be needed and what would their total complexity be? An average of, say, five per protein product would create a total requirement for approximately 400,000 such sequences, which, if they have an average length of 200 bp, would account for a total of 8×10^7 bp. This is still only about 1/20 of the total number of bp involved in functional clusters, although comparable to the number involved in coding for proteins. The requirement for about 400,000 sequences is not inconsistent with the data on the small RNAs, including the family of sequences that share a site for the Alu restriction enzyme [40]. It would, after all, need variation in only nine or 10 positions out of the 200 or more bp to create the required number of different sequences ($4^{10} = 1{,}048{,}576$). Thus, a mechanism of control of expression for all the protein products via the small RNAs is not inconsistent with the existence of a limited number of very large families of these molecules and perhaps a larger number of smaller families. If these speculations turn out to be correct, then it will, of course, be of great interest to know the genetic relationship between these controlling sequences and the gene clusters whose expression they control.

This overall view of the level of complexity of the human genome has, in my opinion, quite fundamental implications. The latest edition of McKusick's catalog [41] already lists a few thousand inherited variants, some common, many, of course, quite rare. If the number of basic genetic functions in terms of gene clusters is really of the order of 5,000 to 10,000, then the proportion of these already represented in the catalog might be quite high. We may, in other words, be much closer to knowing something about a relatively high proportion of the number of basic genetic functions than was realized until quite recently. I am sure that many human geneticists, certainly myself included, have until recently assumed that we have seen only the tip of the mutant iceberg and are nowhere near finding representative mutants for all the basic genetic functions; and this in itself may be quite wrong if genes are clustered and the number of clusters is in the thousands. A major implication is that the chance of finding a gene that affects any particular attribute is no longer negligible, and this, of course, is important, particularly for complex phenotypes such as the chronic diseases, behavior, and physiognomy.

DNA POLYMORPHISMS AND THEIR USES

DNA cloning and sequencing techniques will eventually give us the whole sequence of the human genome. Given the present extraordinary rate of technical advance, and the fact that the problem is not conceptual but one of sheer effort, perhaps we shall see the complete sequence within our lifetime. But whether we shall know what it all means, I very much doubt. Before that extraordinary point is reached, however, much information will be gained from knowledge of specific sequences and, in particular, from the use of restriction enzyme DNA polymorphisms identified by using appropriate clones. This is, of course, the technique that has been pioneered so elegantly for the prenatal detection of hemoglobinopathies by Kan and Dozy [42]. The level of polymorphism so far observed using these techniques, even though the restriction enzymes sample only a small proportion of possible DNA sequences, is extraordinarily high (see, e.g., [38]). These approaches should, therefore, provide us with a quantum jump in the range of available genetic markers for population and family studies. Obviously, if suitable polymorphic markers can be found that are closely linked to inherited diseases such as cystic fibrosis and muscular dystrophy, then these could be used for prenatal diagnosis as in the case of the hemoglobinopathies, without the need for any understanding of the basic functional defects involved. Linked genetic markers will similarly provide the main answer to the unraveling of the genetic determination of complex attributes for which, so far, variance analysis and heritability approaches have provided only a very limited horizon. As pointed out by Solomon and Bodmer [43], a limited range of perhaps some 200 to 300 suitably selected probes might be enough to provide a genetic marker for, say, every 10% recombination. This should be enough, barring statistical artefacts, to identify any genetic component in a complex attribute, even though the functional basis for the attribute is incompletely, if at all, understood.

The principle of this approach is very clearly illustrated by the various HLA and disease associations (see, e.g., [44]). A population association between, say, insulin-

dependent juvenile onset diabetes and HLA-DR3 or DR4 may be due to the effects of the DR3 and DR4 products themselves, or to linkage disequilibrium between these markers and other factors of the HLA system contributing to the disease susceptibility. In families, studies of HLA-linked effects do not depend on linkage disequilibrium and can be used even for traits with no obvious Mendelian basis, so long as enough families can be found with more than one affected individual. By looking at the distribution of HLA haplotypes shared among affected sib pairs, information can be obtained about HLA linkage with a presumed disease susceptibility gene by analogy with the Penrose sib pair method for linkage detection [45]. Since the HLA system is so polymorphic, essentially every multiple-case family is likely to be informative. For example, in the case of insulin-dependent juvenile onset diabetes, a recent summary showed that out of 263 affected sib pairs whose families were typed for HLA, 59% shared two haplotypes, 37% one, and only 4% shared no haplotypes [46]. The expected distribution in the absence of an HLA-linked effect would be 25%, 50%, and 25%, respectively. The difference is highly significant and can be used not only to show the existence of an HLA-linked effect but also to distinguish dominant from recessive inheritance and to estimte disease susceptibility gene frequencies [45, 46].

The simplest strategy for using a comprehensive set of polymorphic markers for studying genetic linkage with a set of discrete attributes, or if necessary with a dichotomized continuous distribution, is to follow the same approach as in the HLA studies, namely, to look for a disturbance of genetic marker segregation among individuals within a family sharing a given phenotypic classification. Clearly, more information could, in principle, be derived from more complete family studies, but the lower the familial concentration of a particular phenotype, the less the information that can be obtained from the complementary class; for example, normals as compared to affecteds in a disease association study. Problems will naturally arise, for example, from associations with several markers, from the logistics of studying the range of markers needed in each family, and from problems of the significance of the extreme deviates that will, of course, be picked up in such a study. These are, however, all surmountable, and there seems no reason to doubt the power of this approach to the analysis of complex phenotypes.

There is a variety of possible approaches to obtaining a suitable set of polymorphic DNA markers. Clearly, one avenue will be through the use of clones obtained for known gene products, as in the case of hemoglobin. Another approach is simply to use random clones from a genomic library, selected to exclude repetitive sequences and to establish their linkage relationship using conventional family studies (see, e.g., [47]). The approach favored in our laboratory, however, is to make use in the first instance of human-mouse somatic cell hybrids to identify clones from defined genetic regions and then to select from these ones that conveniently identify useful polymorphism. Hybrids having single or a limited number of human chromosomes can be used either for the production of, or for the detection of, such clones in a way that is closely analogous to the identification of antigenic differences coded for by particular chromosomes [48, 49]. This approach depends on the differential annealing of human probes to human as compared to mouse DNA and has been

used successfully for the mapping of the hemoglobin genes [50] as well as for the identification of clones from a given chromosome [51, 52]. In our laboratory, we are using a combination of these various approaches to identify DNA clones from particular genetic regions. For example, cDNA clones or genomic clones selected by annealing, under appropriate conditions, with total human DNA to contain only unique sequences, are tested using the Southern blotting technique against combinations of human and mouse cell lines and appropriate human-mouse somatic cell hybrids to localize them to a given chromosome. Finer localization of the position of such clones can be achieved using standard techniques with chromosome breaks [53], or, alternatively, clones can be selected to be from a given limited region, for example, by using cDNA probes from hybrids to pick out appropriate clones from a genomic library [54]. A number of clones obtained from a given region can then be screened for their ability to detect restriction enzyme polymorphisms. There may be a case for being very selective at this stage. For example, it may be easiest to use a limited set of restriction enzymes and choose only those clones that show a suitable level of polymorphism with one or other of a set of conveniently chosen enzymes. This may simplify the logistics of screening with a large number of genetic markers.

Screening may be done initially using a bank of DNA samples obtained from lymphoblastoid cell lines grown out from known individuals. Obviously, such lines can be derived from suitable families, population samples, or other selected individuals and so could provide a bank of DNA to be used for screening and testing of appropriate polymorphisms. Lymphoblastoid cell lines can also provide suitable DNA standards for the identification of polymorphisms by different laboratories.

The task is considerable, but by no means insurmountable, and the rewards enormous. No single laboratory is likely to be in a position to obtain a suitable complete set of genetic markers, and so, undoubtedly, progress in this field will be greatly helped by active collaboration between the various laboratories involved in this work. Following the model of the HLA typers (as documented in the proceedings of the various histocompatibility testing workshops—see, e.g., [55, 56]), comparatively rapid progress could be made toward obtaining a suitable set of genetic markers.

One obvious application of these approaches, in which our laboratory has an interest, is in the detection of the presumed genetic changes that take place in the cancer cell and, of course, also in the detection of genetic markers for familial susceptibility to certain cancers. Using, for example, somatic cell hybrids containing the Philadelphia chromosome, we are aiming to clone the DNA sequences that straddle a specific translocation point and so hope to find out what the significance of such a specific genetic event is to the development of the leukemia. Linked polymorphic markers to DNA repair deficiencies should also, for example, be helpful in establishing to what extent heterozygotes for such deficiencies have an increased susceptibility to cancer [57].

There is one major remaining technical problem in the application of DNA cloning techniques to the analysis of the human genome. This is the gap between the size of the largest DNA clone that can so far be obtained using bacterial systems, which is between 50,000 and at most 100,000 bp, and the smallest cytologically

visible chromosome fragment that can be identified, which must be of the order of 15×10^6 bp. This means that several hundred DNA clones may be needed to cover the smallest visible chromosome fragment, which poses a problem for finding a clone close to a known marker for which there is no known product, and so no direct cloning approach. For example, suppose we have identified a clone that segregates in a human-mouse somatic cell hybrid with the Philadelphia chromosome, it may still take up to a thousand clones to "walk along" the chromosome until we reach the desired translocation point. For this purpose, it would be a great advantage to be able to clone DNA pieces containing of the order of 1,000,000 bp. Perhaps this can be achieved either by cloning such a large piece as a supernumerary chromosome in *E. coli*, for example, using the *E. coli* replication origin itself as a cloning vector or, alternatively, using somatic cells themselves as a means of cloning and introducing chromosome fragments obtained by X-irradiation using chromosome transformation [58]. The major disadvantage of this latter approach is the need for a selective marker on the piece to be cloned, but perhaps the use of transformation for the Herpes simplex thymidine kinase can help in this respect [59, 60].

GENETIC DATA BASES AND FUTURE IMPLICATIONS

Shortly after I went to Oxford and was in the process of having the laboratories remodeled, a sign appeared outside the genetics laboratory saying "Alterations-Department of Genetics." Then, as now, and I imagine for the foreseeable future, it will not be possible to offer desired (or undesired) genetic alterations at will. However, the DNA techniques may indeed soon allow us to *detect* alterations that have occurred or that distinguish one individual from another. This knowledge will still not, however, tell us which of the genetic variations are functionally relevant. A comparable problem at a much simpler level is the difficulty in establishing which amino acid differences between two related protein products may be responsible for a particular serologically detected antigenic determinant.

The challenge of the sequence, although considerable, will surely be solved much more easily than the challenge of the functional meaning of the sequence. We may find, for example, the DNA sequence that straddles a specific translocation point consistently found in a given cancer, but what will it tell us? It may be possible to detect the products made by the sequences on either side of the translocation point either by in vitro translation techniques or simply by reading off the DNA sequences. However, when we have so identified the product, how will we proceed to find out its function and, then, why a particular change is relevant to the progression of the tumor?

Information will undoubtedly be accumulated slowly and steadily about the 5,000 or so clusters of gene products and their functions. The more we know, the easier it will be to establish the likely functional significance of a given genetic difference. Already, it seems now that a systematic collection of the information on all known genes and gene products, namely, all proteins whether structural, enzyme, or hormone, could be of value and help to interrelate functionally known products and serendipitously identified DNA sequences. Eventually, when we have the complete genome sequence library, having found a linked marker for some part

of a complex phenotype, we may simply be able to look up what are the gene clusters in its neighborhood and through that focus onto the functional basis of the phenotype. The fact that there may be a relatively limited number of basic genetic functions, together with knowledge about their interrelationships, should eventually make it possible to discern the functional and structural basis for any inherited differences, however complex. I have no doubt that in due course we shall be able to answer the question of why little Johnny looks like his mother.

I am sure many of you must have wondered, as I have, about the significance of the apparent correlation between facial features and behavioral characteristics, an association expressed by the word physiognomy, defined by the *Oxford English Dictionary* as "face as an index of character: art of judging character from face and form." Is this association an illusion? Does the character mould the face, or is there really a true biological association between face and behavior? In the latter case, this must surely mean that the polymorphic genetic markers determining these various attributes remain associated at the population level, since it is hard to see how polymorphic differences in a single gene cluster could explain both sets of attributes, namely, face and behavior. In that case, population association implies linkage disequilibrium which in turn implies close linkage. Are, therefore, the genes that control facial features and certain aspects of behavior really closely linked to each other? The DNA sequence will no doubt provide the answer in due course.

The revolution in biology that forms the main theme for this paper came originally from the introduction of both the physical sciences and physical scientists into biology. Perhaps the revolution that is surely needed in sociology and economics to improve the management of our complex modern society will come from the contributions of biology and biologists to these areas. The whole DNA sequence will eventually be known, and also, but even more eventually, its meaning will be understood. This knowledge will have profound implications for all aspects of human activities and endeavors and surely will, in the long run, contribute positively to the betterment of our society.

REFERENCES

1. Perutz M: Molecular biology of the future. *New Scientist* Jan. 31, 1980
2. Barnstable CJ, Jones EA, Bodmer WF: Genetic structure of major histocompatibility regions, in *International Review of Biochemistry. Defence and Recognition IIA, 22, Cellular Aspects*, vol 4, Baltimore, University Park Press, 1979, pp 151–225
3. Darlington CD, Mather K: *The Elements of Genetics*. London, Allen and Unwin, 1949
4. Bodmer WF: HLA structure and function: a contemporary view. *Tissue Antigens* 17:9–20, 1981
5. Barnstable CJ, Jones EA, Crumpton MJ: Isolation, structure and genetics of HLA-A, -B, -C and -DRw (Ia) antigens. *Br Med Bull* 34:241–266, 1978
6. Winchester RJ, Kunkel HG: The human Ia system. *Adv Immunol* 28:222–292, 1979
7. Bodmer WF: The major histocompatibility gene clusters of man and mouse, in *Mammalian Genetics and Cancer: The Jackson Laboratory 50th Anniversary Symposium*, edited by Russell ES, New York, Alan R. Liss, 1981, pp 213–240
8. Jones PP, Murphy DB, McDevitt HO: Two-gene control of the expression of a murine Ia antigen. *J Exp Med* 148:925–939, 1978
9. Barnstable CJ, Jones EA, Bodmer WF, et al.: Genetics and serology of HLA linked human Ia antigens. *Cold Spring Harbor Symp Quant Biol* 41:443–455, 1977

10. LEE JS, TROWSDALE J, BODMER WF: Synthesis of HLA antigens from membrane-associated messenger RNA. *J Exp Med* 152:3–10, 1980
11. BODMER WF: Models and mechanisms for HLA and disease associations. *J Exp Med* 152:353–357, 1980
12. BODMER WF: Evolutionary significance of the HL-A system. *Nature* 237:139–145, 1972
13. BODMER WF, BODMER JG: Evolution and function of the HLA system. *Br Med Bull* 34:309–316, 1978
14. BODMER WF, ED.: The HLA system. *Br Med Bull* 34:213–316, 1978
15. IVANYI P: Some aspects of the H-2 system, the major histocompatibility system in the mouse. *Proc R Soc Lond* 202:117–158, 1978
16. FRANCKE U, WEITKAMP LR: Report of the committee on the genetic constitution of chromosome 6. *Cytogenet Cell Genet* 25:32–38, 1979
17. CRICK F: Split genes and RNA splicing. *Science* 204:264–271, 1979
18. PONTECORVO G: *Trends in Genetic Analysis*. New York, Columbia Univ. Press, 1959
19. BRODSKY FM, PARHAM P, BARNSTABLE CJ, CRUMPTON MJ, BODMER WF: Monoclonal antibodies for analysis of the HLA system. *Immunol Rev* 47:1–61, 1979
20. KRANGEL MS, ORR HT, STROMINGER JL: Structure, function and biosynthesis of the major human histocompatibility antigens (HLA-A and HLA-B). *Scand J Immunol* 11:561–571, 1980
21. COLIGAN JE, KINDT TJ, EWENSTEIN BM, UEHARA H, NISIZAWA T, NATHANSON SG: Primary structure of murine MHC alloantigens. II. Amino acid sequence studies of the cyanogen bromide fragments of the H-2K^b glycoprotein. *Proc Natl Acad Sci USA* 75:3390–3394, 1978
22. DEMANT P, IVANYI D, NEAUPORT-SAUTES C, SNOEX M: H-2.28, an alloantigenic marker allelic to H-2.1, is expressed on all three known types of H-2 molecules. *Proc Natl Acad Sci USA* 75:4441–4445, 1978
23. BODMER WF: A new genetic model for allelism at histocompatibility and other complex loci: polymorphism for control of gene expression. *Transplant Proc* 4:1471–1475, 1973
24. LERNER MR, BOYLE JA, MOUNT SM, WOLIN SL, STEITZ JA: Are snRNPs involved in splicing? *Nature* 283:220–224, 1980
25. ROGERS J, WALL R: A mechanism for RNA splicing. *Proc Natl Acad Sci USA* 77:1877–1879, 1980
26. BRIDGES CB: Duplication. *Anat Rec* 15:357–358, 1919
27. HOROWITZ MH: On the evolution of biochemical synthesis. *Proc Natl Acad Sci USA* 31:153–157, 1945
28. LEWIS EB: Pseudoallelism and gene evolution. *Cold Spring Harbor Symp Quant Biol* 16:159–174, 1951
29. WEATHERALL DJ, CLEGG JB, WOOD WG, PASVOL G: Human haemoglobin genetics. Human genetics: possibilities and realities. *Ciba Found Symp* 66:147–186, 1979
30. PORTER RR, REID KBM: Activation of the complement system by antibody-antigen complexes: the classical pathway. *Adv Protein Chem* 33:1–71, 1979
31. BODMER WF: Gene clusters and the HLA system. Human genetics: possibilities and realities. *Ciba Found Symp* 66:205–229, 1979
32. FIDDES JC, GOODMAN HM: The cDNA for the β-subunit of human chorionic gonadotropin suggests evolution of a gene by readthrough into the 3′-untranslated region. *Nature* 286:684–687, 1980
33. BODMER WF: Molecular and genetic organisation: the future. Human genetics: possibilities and realities. *Ciba Found Symp* 66:395–400, 1979
34. NISHIOKA Y, LEDER A, LEDER P: Unusual α-globin-like gene that has cleanly lost both globin intervening sequences. *Proc Natl Acad Sci USA* 77:2806–2809, 1980
35. VAN DER BERG J, VAN OOYEN A, MANTEI N, ET AL.: Comparison of cloned rabbit and mouse β-globin genes showing strong evolutionary divergence of two homologous pairs of introns. *Nature* 276:37–44, 1978

36. GUTZ H, LESLIE JF: Gene conversion: a hitherto overlooked parameter in population genetics. *Genetics* 83:861–866, 1976
37. WALKER PMB: Genes and non-coding DNA sequences. Discussion. Human genetics: possibilities and realities. *Ciba Found Symp* 66:41–42, 1979
38. JEFFREYS AJ: DNA sequence variants in the γ^{G}-, γ^{A}-, δ- and β-globin genes of man. *Cell* 18:1–10, 1979
39. ROYAL A, GARAPIN A, CAMI B, ET AL.: The ovalbumin gene region: common features in the organisation of three genes expressed in chicken oviduct under hormonal control. *Nature* 279:126–132, 1979
40. RUBIN CM, HOUCK CM, DEININGER PL, FRIEDMANN T, SCHMID CW: Partial nucleotide sequence of the 300-nucleotide interspersed repeated human DNA sequences. *Nature* 284:372–374, 1980
41. MCKUSICK VA: *Mendelian Inheritance in Man*, 5th ed. Baltimore, Johns Hopkins Univ. Press, 1978
42. KAN YW, DOZY AM: Polymorphism of DNA sequence adjacent to human β-globin structural gene: relationship to sickle mutation. *Proc Natl Acad Sci USA* 75:5631–5635, 1978
43. BODMER WF, SOLOMON E: Evolution of sickle variant gene. *Lancet* April 28:923, 1979
44. BODMER WF: The HLA system and disease. The Oliver Sharpey Lecture 1979. *J R Coll Physicians Lond* 14:43–50, 1980
45. THOMSON G, BODMER WF: The genetic analysis of HLA and disease associations, in *HLA and Disease*, edited by DAUSSET J, SVEJGAARD A, Copenhagen, Munksgaard, 1980, pp 84–93
46. SVEJGAARD A, PLATZ P, RYDER LP: Insulin-dependent diabetes mellitus. Joint results of the 8th workshop study, in *Histocompatibility Testing 1980*, edited by TERASAKI PI, Los Angeles, Univ. of California Press, 1980
47. BOTSTEIN D, WHITE RL, SKOLNICK M, DAVIS RW: Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet* 32:314–331, 1980
48. BUCK DW, BODMER WF: The human species antigen on chromosome 11. Proceedings of the 2nd International Workshop on Human Gene Mapping. *Cytogenet Cell Genet* 14:87–89, 1975
49. BARNSTABLE CJ, BODMER WF, BROWN G, ET AL.: Production of monoclonal antibodies to group A erythrocytes HLA and other human cell surface antigens—new tools for genetic analysis. *Cell* 14:9–20, 1978
50. JEFFREYS A, CRAIG IW, FRANCKE U: Localisation of the γ^{G}-, γ^{A}-, δ - and β-globin genes on the short arm of human chromosome 11. *Nature* 281:606–608, 1979
51. GUSELLA JF, KEYS C, VARSANYI-BREINER A, ET AL.: Isolation and localization of DNA segments from specific human chromosomes. *Proc Natl Acad Sci USA* 77:2829–2833, 1980
52. WOLF SF, MARENI CE, MIGEON BR: Isolation and characterisation of cloned DNA sequences that hybridise to the human X chromosome. *Cell* 21:95–102, 1980
53. GOSS SJ, HARRIS H: New method for mapping genes in human chromosomes. *Nature* 255:680–684, 1975
54. MANIATIS T, HARDISON RC, LACY E, ET AL.: The isolation of structural genes from libraries of eucaryotic DNA. *Cell* 15:687–701, 1978
55. BODMER WF, BATCHELOR JR, BODMER JG, FESTENSTEIN H, MORRIS PJ, EDS.: *Histocompatibility Testing 1977: Report 7th International Histocompatibility Testing Workshop Conference*. Copenhagen, Munksgaard, 1978
56. TERASAKI PI, ED.: *Histocompatibility Testing 1980*. Los Angeles, Univ. of California Press, 1980
57. POLANI PE: DNA repair defects and chromosome instability disorders. Human genetics: possibilities and realities. *Ciba Found Symp* 66:81–133, 1979

58. McBride OW, Ozer HL: Transfer of genetic information by purified metaphase chromosomes. *Proc Natl Acad Sci USA* 70:1258–1262, 1973
59. Munyon W, Kraiselburd E, Davis D, Mann J: Transfer of thymidine kinase to thymidine kinaseless L cells by infection with ultraviolet-irradiated herpes simplex virus. *J Virol* 7:813–820, 1971
60. Wigler M, Silverstein S, Lee LS, Pellicer A, Cheng YC, Axel R: Transfer of purified Herpes virus thymidine kinase gene to cultured mouse cells. *Cell* 11:223–232, 1977